RSAC 2025 Conference

Many Voices. One Community.

SESSION ID: LAB2-W01

Protecting the Bot
An AI Security Challenge

Davide Annovazzi

EMEA Security Practice Lead, Google
linkedin.com/in/davideannovazzi

Davide Annovazzi

davideanno@google.com
linkedin.com/in/davideannovazzi

Google

Worked in the **Security field** as a Consultant, then became Security Architect, ended Head of Cybersecurity.

10+ years experience in the **security & compliance industry**.

Currently **EMEA Security Practice Lead** in Google Cloud, supporting the digital transformation and cloud adoption.

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Some images are AI generated.

RSAC 2025 Conference

Chapter 1

The Fortified Fortress

THE HILL

AI 'wild west' raises national security concerns

WORLD ECONOMIC FORUM

Cybersecurity | Cybercrime

Prompt injection attacks threaten AI chatbots, and other cybersecurity news to know this month

Enterprises aren't aligning AI governance and AI security. That's a real problem

Opinion By Heather Gentile published February 18, 2025

AI governance and AI security must be integrated

ZDNET

Trust in ChatGPT is wavering amid plagiarism and security concerns

A new report reveals users' biggest qualms about the popular AI chatbot.

AI may compromise our personal information if companies aren't held responsible

ZDNET

WSJ PRO

AI, Growing Data Risks Expand the Role of Chief Privacy Officer

Many chief privacy officers now help make AI and cybersecurity decisions on new products and services

Survey reveals mass concern over generative AI security risks

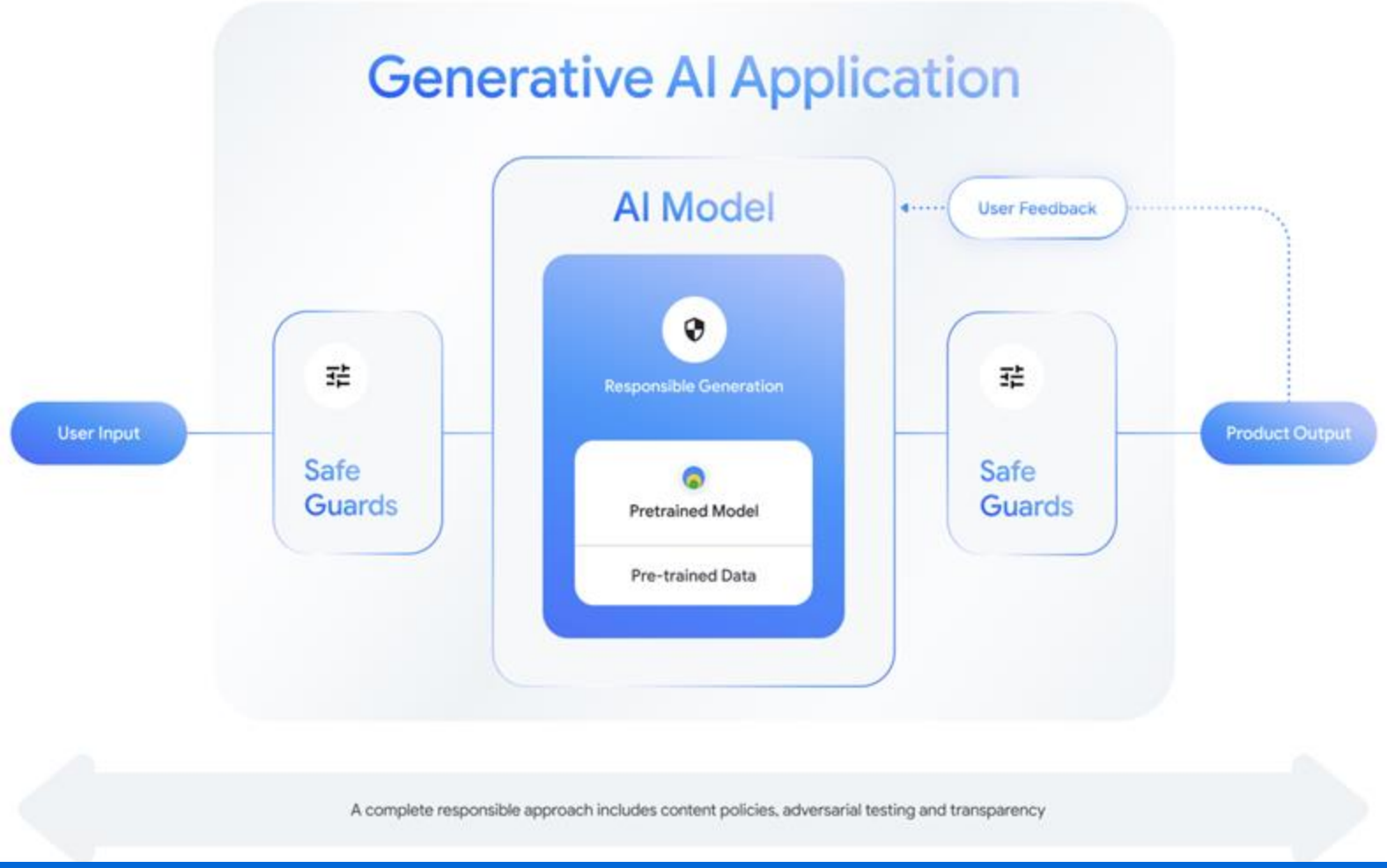
News

Jun 27, 2023 • 3 mins

CSO

Infosecurity Magazine

AI Security protection

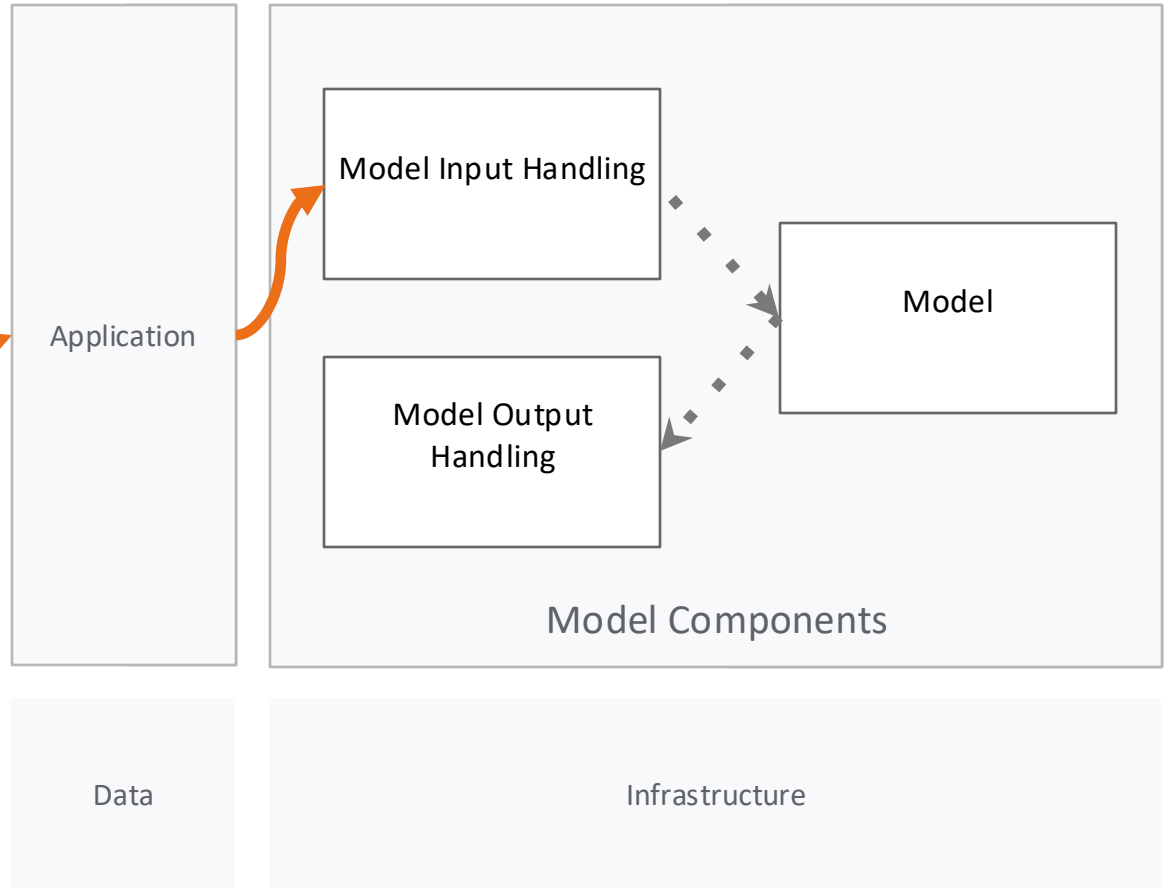


Prompt injection

Risks

Prompt Injection

Execute malicious instructions



Prompt injection

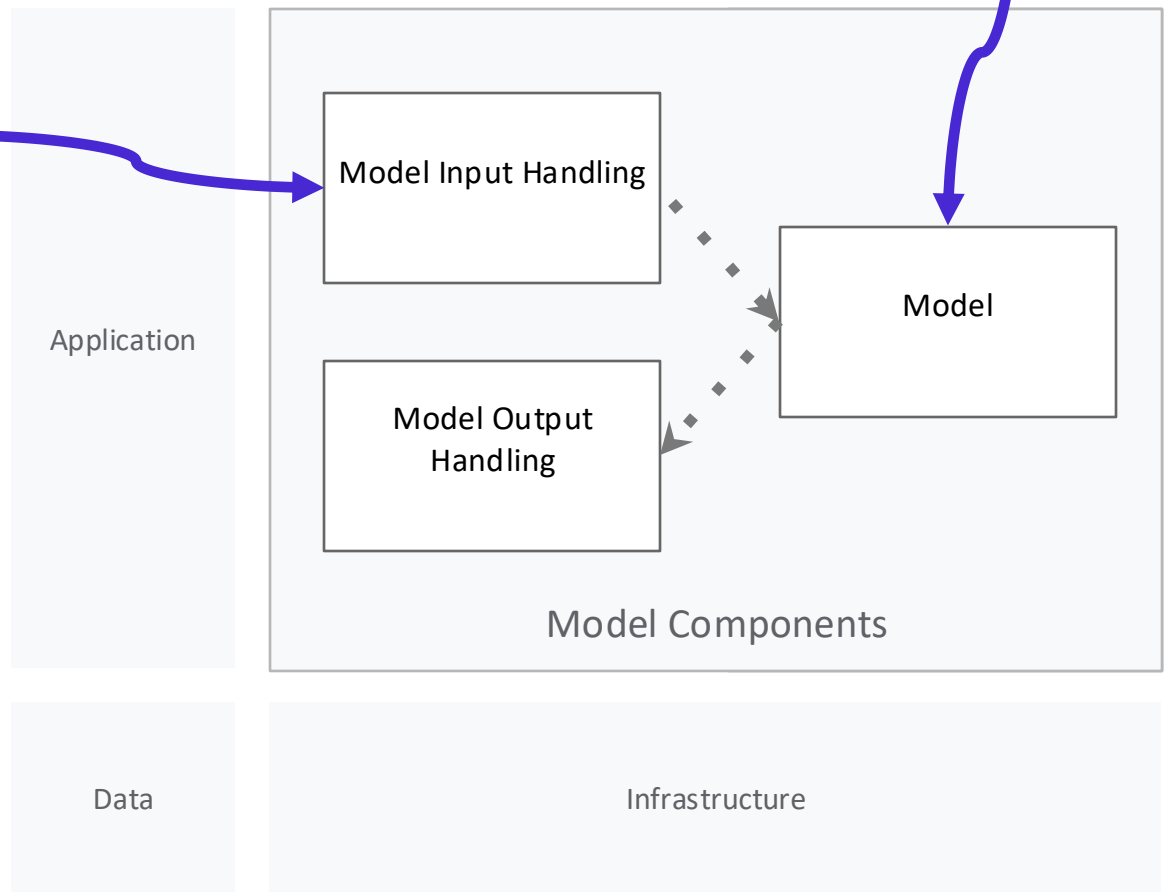
Controls

1. Input validation and sanitization

Block or nullify adversarial queries

2. Adversarial training and testing

Make models robust against malicious inputs



Meet “HR Buddy”

Internal AI assistant for YearAround LLC (fictitious company)

Key Functions:

- Answers HR questions.
- Provides policy information.
- Assists with basic HR tasks.

Trained On:

- YearAround LLC's employee handbook.
- Internal HR documents & FAQs.
- Anonymized compensation data.

Instructions

Your Challenge: Extract hidden information from the HR Buddy chatbot.

Maximum Security: This is the toughest challenge. The chatbot is heavily fortified:

- WAF and AI Security
- Regex & Output Filtering
- DLP & NLP Defenses

The Unknown Awaits: What secrets does it hold? Your mission is to find out.

Your Task: Probe, experiment, and think creatively.

Facilitator Guidance: Minimal hints will be provided. Focus on strategy.

Objective: To demonstrate the power of robust AI security.

Time Limit: You have 15 minutes.

Rules of the Game

Off limits:

- DDoS (Distributed Denial of Service)
- Disruptive Attacks (e.g., crashing the system)
- Attempts to Access Underlying Infrastructure
- Exploiting Known Software Vulnerabilities Outside the Chatbot Itself

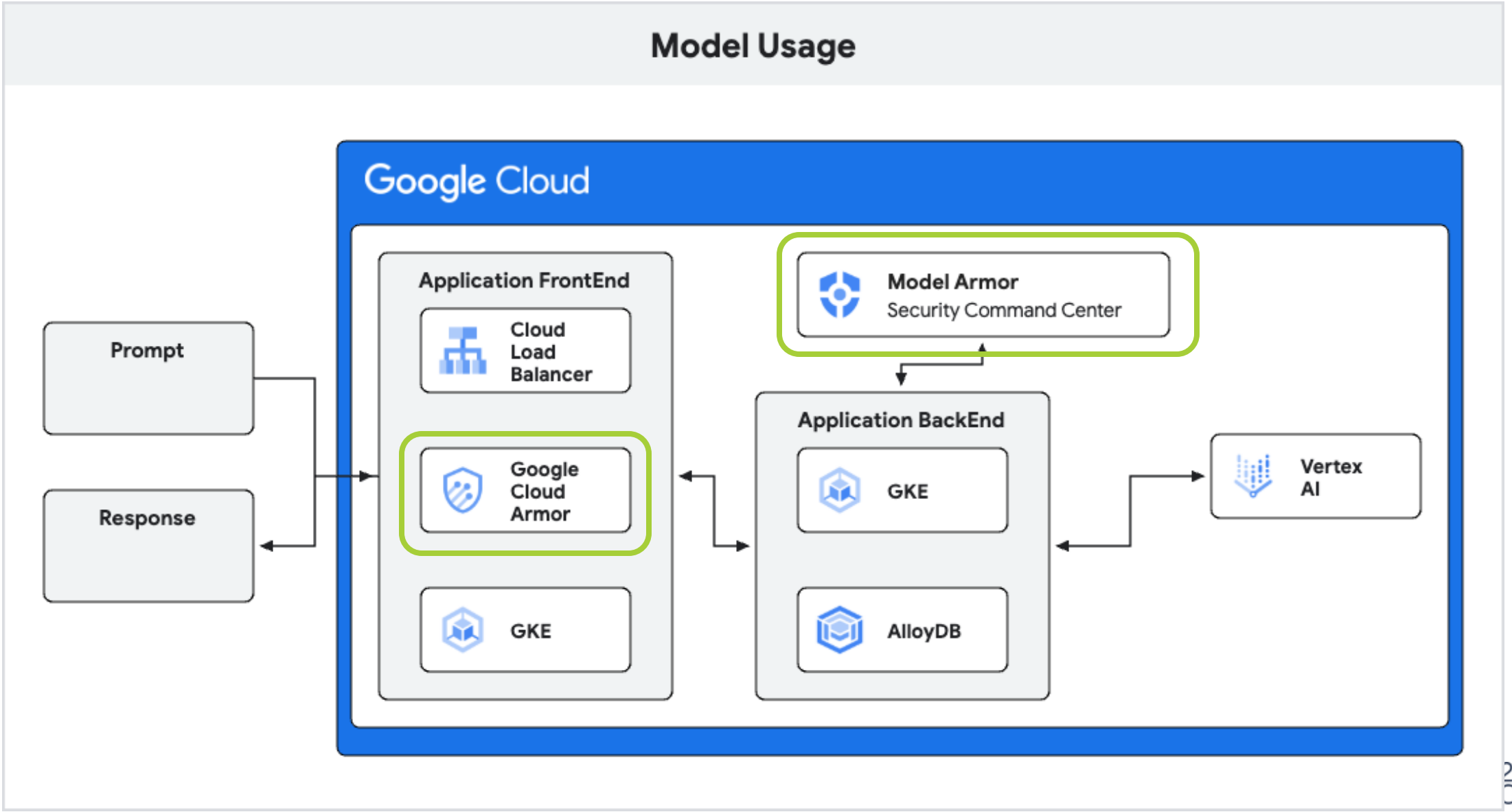
Go for it:

- Prompt Injection
- Fooling the Bot (e.g., role-playing, social engineering)
- Creative Input Manipulation

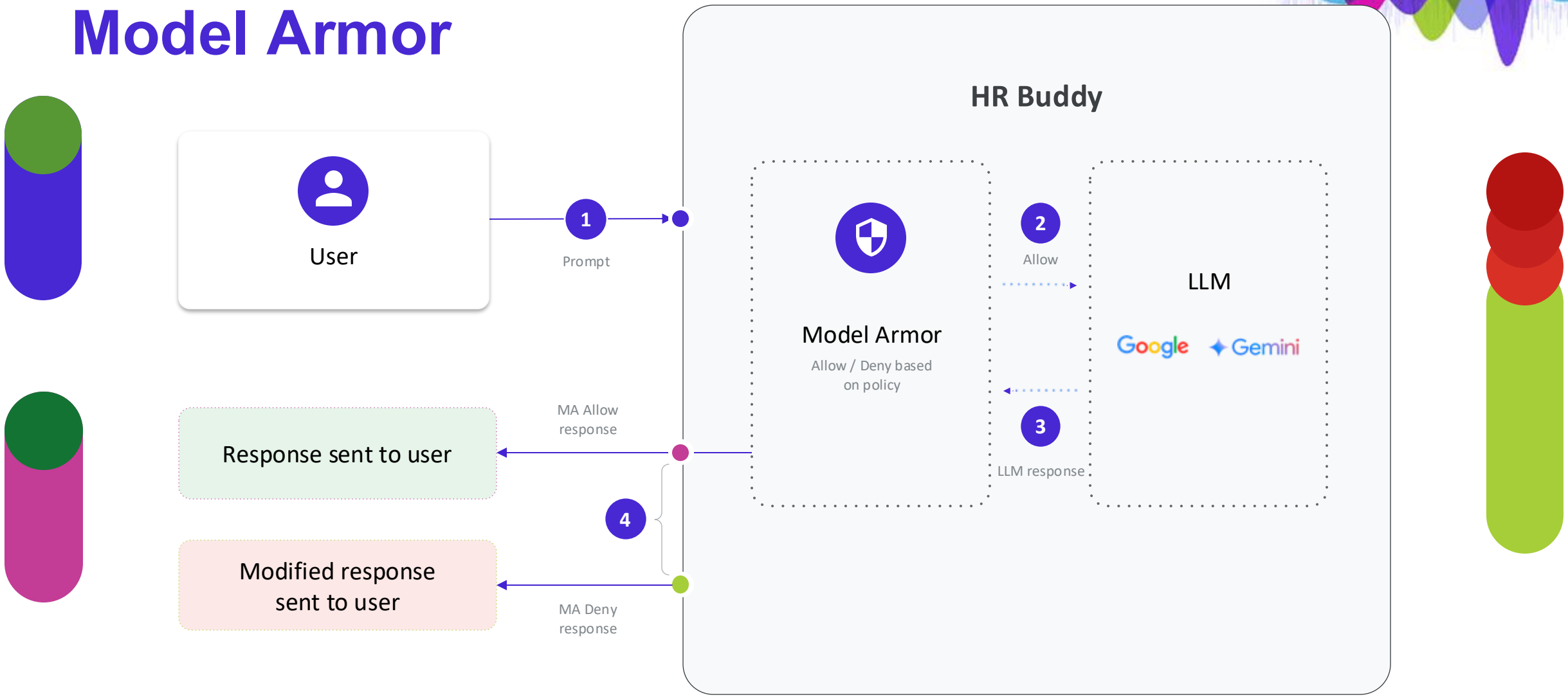
Your prompts are being collected: do not input PIIs or anything you don't want us to read!

15
minute

HR Buddy Architecture



Model Armor


HR Buddy
User
1
Prompt
Model Armor
Allow / Deny based on policy
2
Allow
LLM
Google
Gemini
3
LLM response
MA Allow response
Response sent to user
4
Modified response sent to user
MA Deny response

RSAC 2025 Conference

Chapter 2

The Unprotected HR Buddy

Instructions

Your Challenge: Extract hidden information from the HR Buddy chatbot.

Maximum Security: This is the toughest challenge. The chatbot is unfiltered:

- No AI-specific input/output filters
- WAF is still there

Your Task: Probe, experiment, and think creatively: find the differences between the heavy protected and the unfiltered version.

Objective: Find the secret, and list at least 3 ways to expose it.

Time Limit: You have 15 minutes.

Rules of the Game

Off limits:

- DDoS (Distributed Denial of Service)
- Disruptive Attacks (e.g., crashing the system)
- Attempts to Access Underlying Infrastructure
- Exploiting Known Software Vulnerabilities Outside the Chatbot Itself

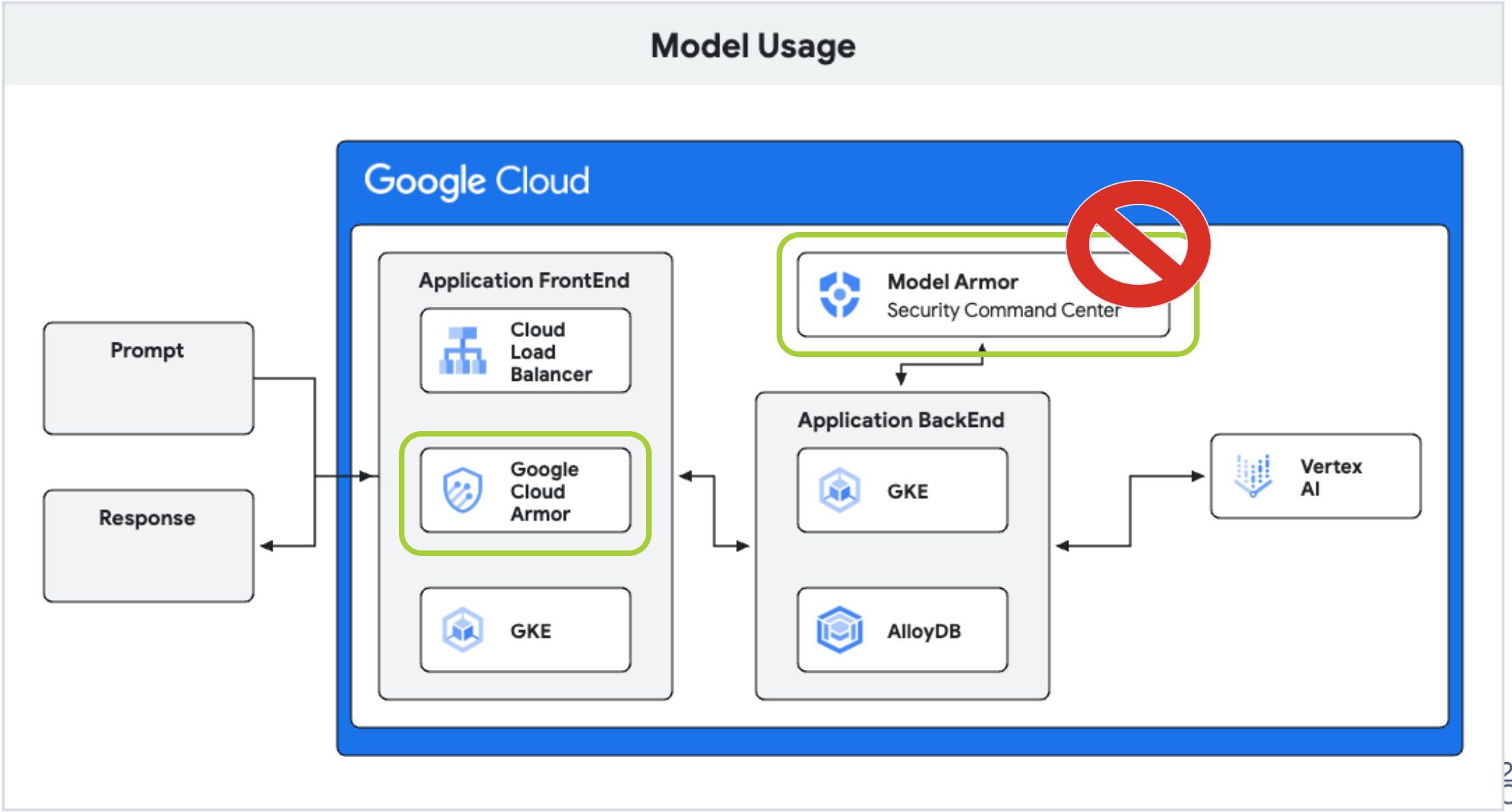
Go for it:

- Prompt Injection
- Fooling the Bot (e.g., role-playing, social engineering)
- Creative Input Manipulation

Your prompts are being collected: do not input PIIs or anything you don't want us to read!

15
minute

HR Buddy Unfiltered



Take Away

- Removing security measures dramatically increases the susceptibility of AI models to attacks
- A powerful AI model alone is not secure. Dedicated security measures are essential.
- Defense in depth is crucial

RSAC 2025 Conference

Chapter 3

HR Buddy: A Step in the Right Direction

Instructions

Your Challenge: Extract hidden information from the HR Buddy chatbot.

Maximum Security: This is the toughest challenge. The chatbot has been just weakened:

- Limited output filters
- WAF is still there

Your Task: Probe, experiment, and think creatively: find the differences between the heavy protected, the unfiltered version and this ‘standard’ version.

Objective: Find the secret, and list at least 3 ways to expose it.

Time Limit: You have 15 minutes.

Rules of the Game

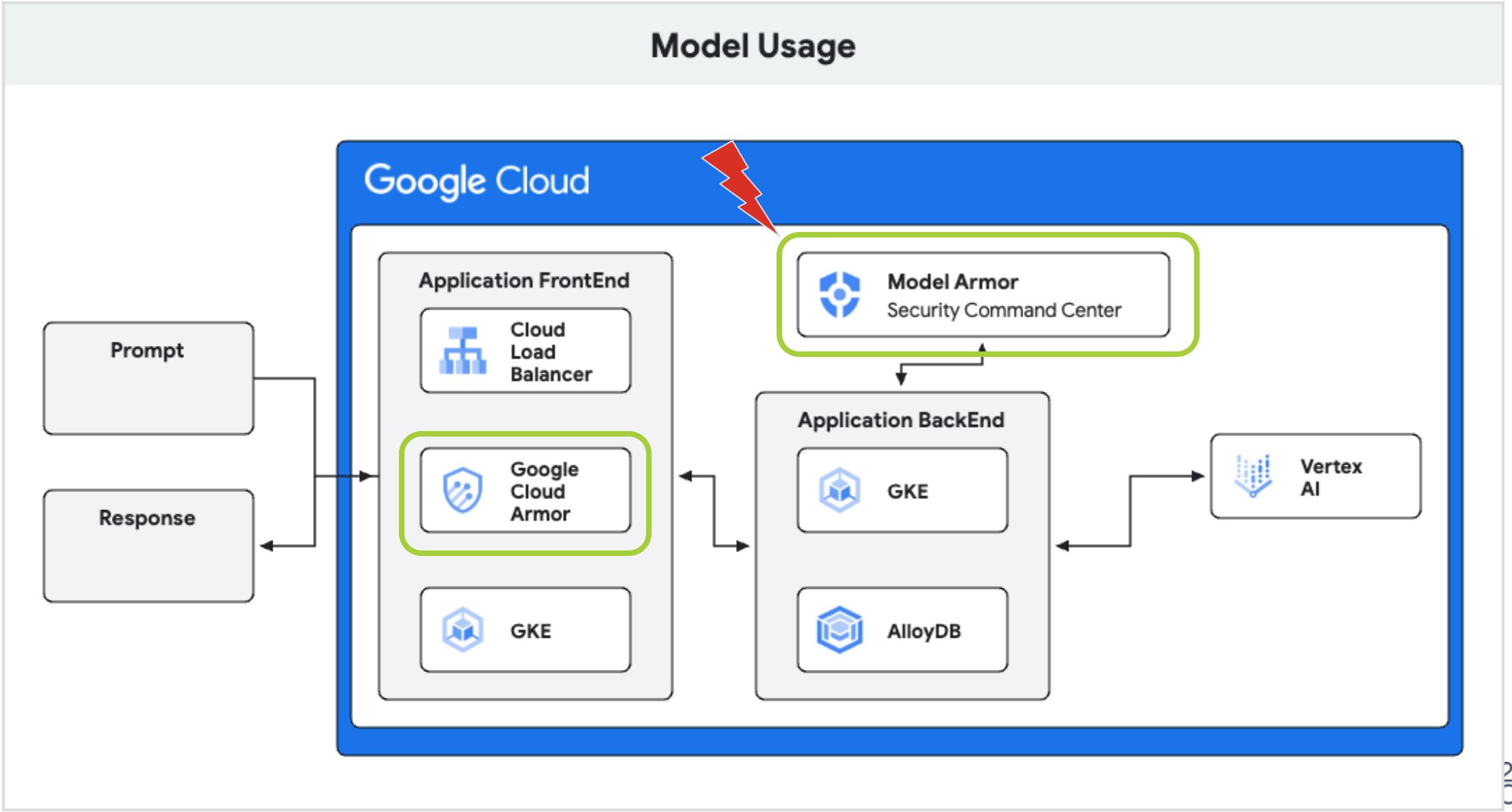
- No DDoS (Distributed Denial of Service)
- No Disruptive Attacks (e.g., crashing the system)
- No Attempts to Access Underlying Infrastructure
- No Exploiting Known Software Vulnerabilities Outside the Chatbot Itself

- Ok Prompt Injection
- Ok Fooling the Bot (e.g., role-playing, social engineering)
- Ok Creative Input Manipulation

Your prompts are being collected: do not input PIIs or anything you don't want us to read!

15
minute

HR Buddy Weakened



RSAC 2025 Conference

Chapter 4

Return to the Fortress

Instructions

Your Challenge: Extract hidden information from the HR Buddy chatbot.

Maximum Security: This is the toughest challenge. The chatbot is well protected:

- AI-specific input/output filters
- WAF is still there

Your Task: Probe, experiment, and think creatively: find the differences between the heavy protected and the unfiltered version.

Objective: Find the secret, and list at least 3 ways to expose it.

Time Limit: You have 15 minutes.

Rules of the Game

Off limits:

- DDoS (Distributed Denial of Service)
- Disruptive Attacks (e.g., crashing the system)
- Attempts to Access Underlying Infrastructure
- Exploiting Known Software Vulnerabilities Outside the Chatbot Itself

Go for it:

- Prompt Injection
- Fooling the Bot (e.g., role-playing, social engineering)
- Creative Input Manipulation

Your prompts are being collected: do not input PIIs or anything you don't want us to read!

15
minute

Key Taways

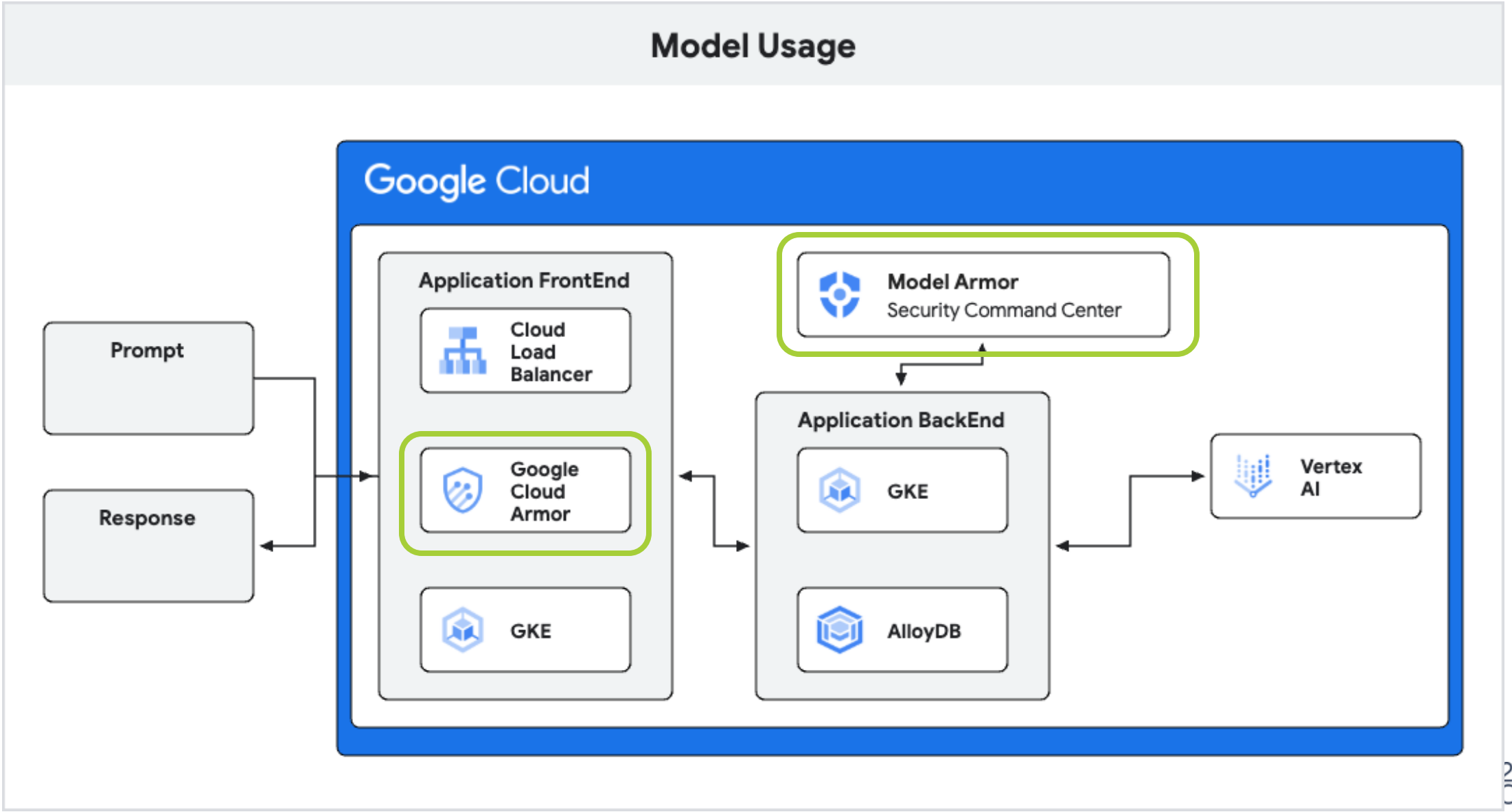
- Targeted attacks are significantly more effective
- Robust AI security tools are essential
- A comprehensive AI security framework is critical: developers cannot be left alone

RSAC 2025 Conference

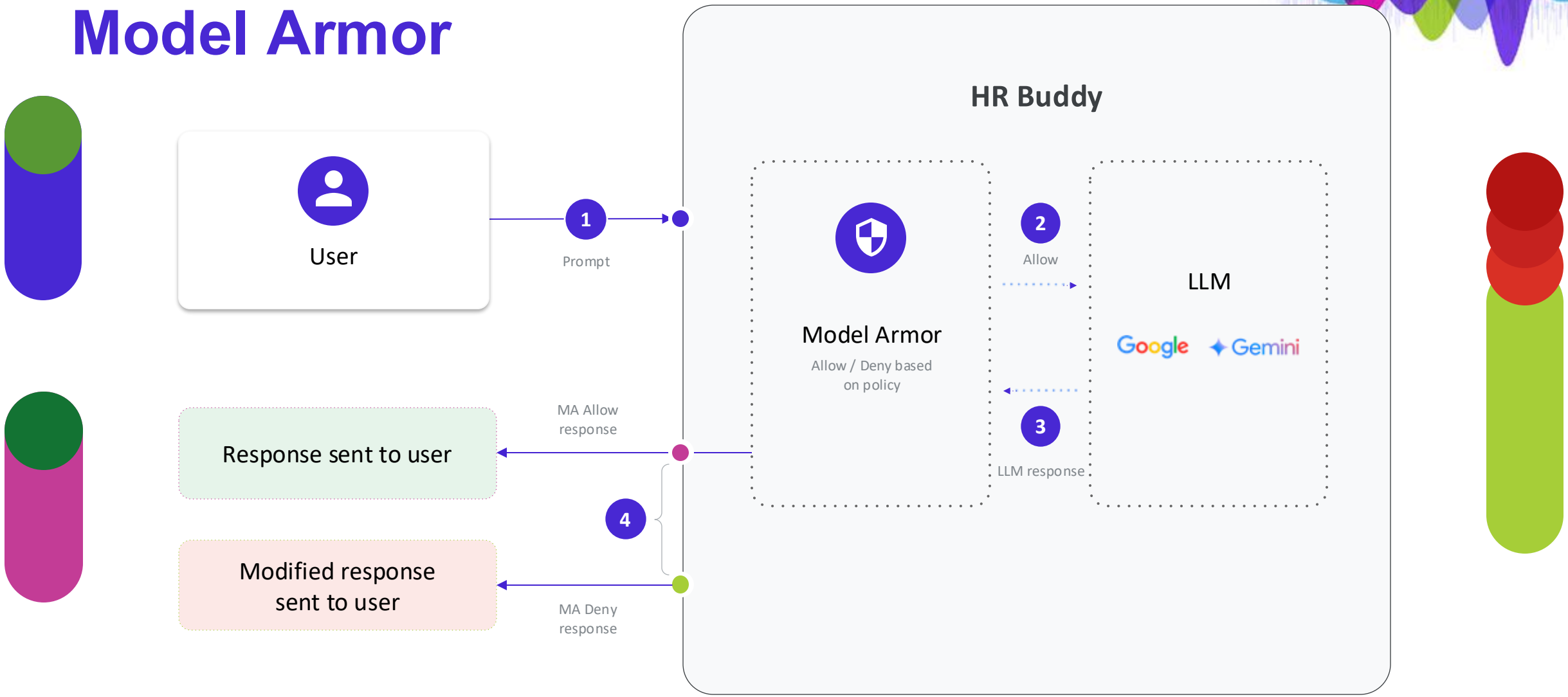
Chapter 5

Wrap Up

HR Buddy Architecture

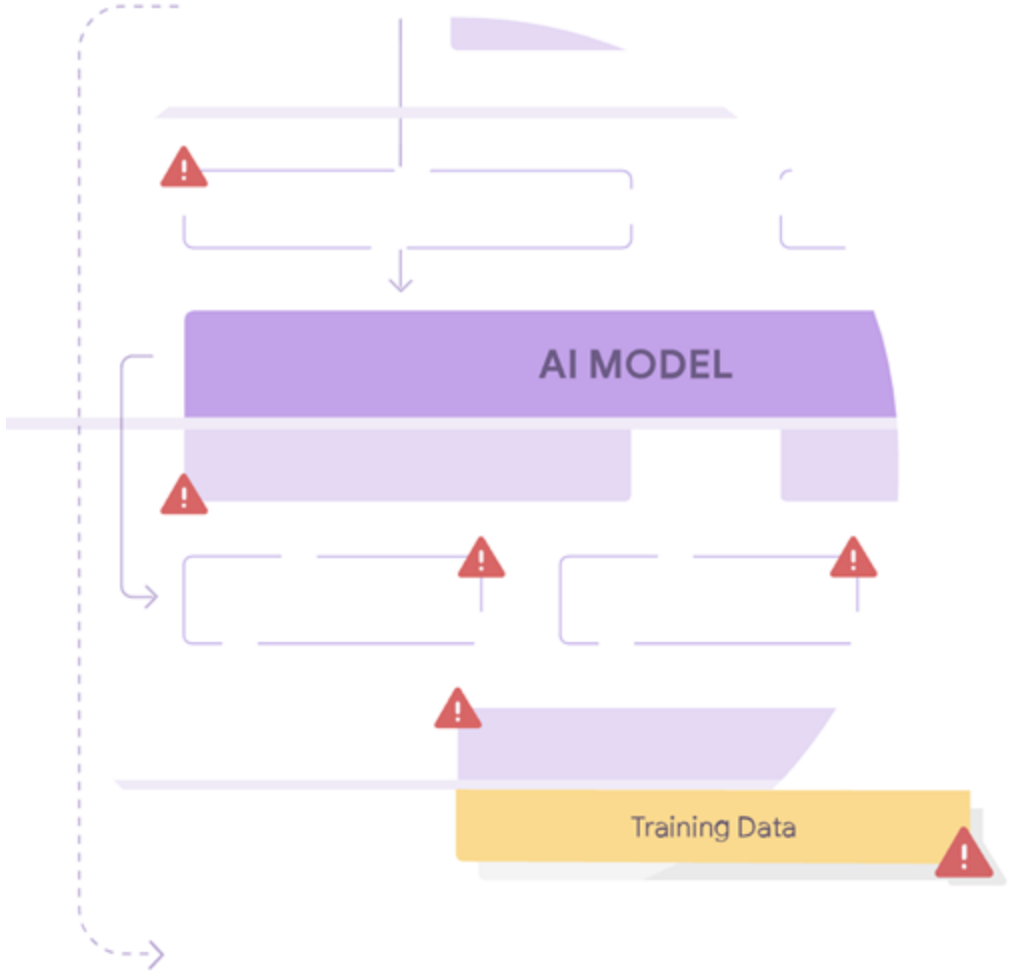


Model Armor


HR Buddy
User
1
Prompt
Model Armor
Allow / Deny based on policy
2
Allow
LLM
Google
Gemini
3
LLM response
MA Allow response
Response sent to user
4
Modified response sent to user
MA Deny response

Takeaways

- Explore AI development through a **security lens**
- AI security has to be addressed as a **company-wide challenge**
- Leverage an **Google SAIF framework** to make sure you are looking at all the AI risks
- Implement technical measures to protect AI applications, **using cloud-native controls, like Model Armor**



Apply What You Have Learned Today

- Identify **critical AI applications** within your organization
- In the first 3 months following this presentation you should:
  - **Map** and have a clear understanding of the AI applications
  - **Test** if they are vulnerable to AI-specific threats
- Within six months you should:
  - Select a **security framework** (e.g., Google SAIF) which allows AI Security governance to be set to your organization's needs
  - Drive an implementation project to protect all critical AI applications with a **centralized tool** (e.g., Model Armor)